Model averaging fails to improve the extrapolation capability of the island species–area relationship

Matthews, Thomas; Aspin, Thomas

DOI:
10.1111/jbi.13598

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Matthews, T & Aspin, T 2019, 'Model averaging fails to improve the extrapolation capability of the island species–area relationship', *Journal of Biogeography*, vol. 46, no. 7, pp. 1558-1568. https://doi.org/10.1111/jbi.13598

Submission to: Journal of Biogeography

Article Type: Research Paper

Model averaging fails to improve the extrapolation capability of the island species–area relationship

Thomas J. Matthews[1, 2] & Thomas W. H. Aspin[1]

[1]GEES (School of Geography, Earth and Environmental Sciences) and the Birmingham Institute of Forest Research, the University of Birmingham, Birmingham, B15 2TT

[2]CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Univ. dos Açores – Depto de Ciências e Engenharia do Ambiente, PT-9700-042, Angra do Heroísmo, Açores, Portugal.

*Correspondence: Thomas J. Matthews, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

Email: t.j.matthews@bham.ac.uk

Running header: ISAR extrapolation

Word count: abstract: 352 words; main text = 5742 words; 2 Tables; 2 Figures; 39 references; 2 appendices

Keywords extrapolation, habitat islands, multi-model inference, model averaging, power model, species–area relationship, species richness

ABSTRACT

Aim: One of the main applications of the island species–area relationship (SAR) is to predict species richness in areas of habitat too large to be sampled, but there are few clear guidelines for choosing an appropriate model for this purpose. We therefore aimed to test whether a multi-model averaging approach could improve the accuracy of predictions made by extrapolating the ISAR. Specifically, we compared the performance of multi-model averaging with that of the default ISAR model of choice, the power model, in predicting species richness in large habitat islands.

Location: Global

Taxa: Vertebrates, invertebrates and plants

Methods: We removed the largest islands from 120 habitat island datasets, and fitted both the power model and a multi-model average curve (averaging the predictions of up to 20 ISAR models) to this filtered dataset. We then assessed the accuracy of both approaches in predicting the species richness of the largest island in the original dataset using the log error of extrapolation (LEE) metric. A generalized additive regression modelling framework was

used to determine whether any dataset characteristics could explain variation in the LEE values for the power model (LEE-POW).

Results: The power model gave the more accurate richness predictions for 58% of the analysed datasets and the multi-model averaged curve gave the more accurate predictions for the remaining 42%. Both the power models (61% of LEE-POW values were positive) and the multi-model averaged curve (60% were positive) had a slightly greater tendency to over predict the observed richness. The confidence intervals were also on average narrower for the power model predictions (median 95% confidence interval width = 18 species) than for the multi-model averaged curve predictions (median 95% confidence interval width = 78). The range in island areas and richness values explained a small amount of the variation in LEE-POW.

Main conclusions: Contrary to expectation, multi-model averaging was less accurate than the power model in the majority of cases, and thus does not appear to be a panacea for uncertainty in model choice when extrapolating the ISAR. However, further research is urgently needed to evaluate the performance of a multi-model averaging approach at larger spatial scales.

INTRODUCTION

The species–area relationship (SAR) describes the near-universally observed pattern whereby the number of species increases with the area sampled (Rosenzweig, 1995; Tjørve & Tjørve, 2017). A number of different types of SARs have been described (Scheiner, 2003; Whittaker & Fernández-Palacios, 2007), and these can be broadly split into island species–area relationships (ISARs), whereby the number of species occurring within each of a set of islands is analysed as a function of the area of each island, and species accumulation curves, which describe the relationship between increasing cumulative species number with increasing sampling area (see Matthews, Triantis, Rigal, Borregaard, Guilhaumon & Whittaker, 2016). This paper is focused on ISARs (Type IV SARs in Scheiner's 2003 typology). Although over twenty ISAR models have been proposed (Tjørve, 2003; Triantis, Guilhaumon & Whittaker, 2012), the most widely used is the power model, $S = c * A^z$, where S is the number of species on an island, A is the area of an island, and *c* and *z* are fitted constants (Arrhenius, 1921). In comparative analyses, the power model has been found to provide the best fit to a number of true and habitat island datasets, but it is not universally the best model (Dengler, 2009; Triantis et al., 2012; Matthews, Guilhaumon, Triantis, Borregaard & Whittaker, 2016), and the ISAR has been found to exhibit forms that the predominantly convex power model cannot provide a good fit to, such as sigmoidal shaped relationships (Lomolino, 2000; Triantis et al., 2012). For example, in an analysis of 182 habitat island datasets, the power model provided the best fit, out of twenty candidate ISAR models, in only 24% of cases (Matthews, Guilhaumon et al., 2016). Put another way, there is considerable model uncertainty in regards to the form of the ISAR, and a number of studies have argued that ISAR analyses should incorporate a wider set of models rather than simply the power model (Guilhaumon, Gimenez, Gaston & Mouillot, 2008; Guilhaumon, Mouillot & Gimenez, 2010; Triantis et al., 2012; Benchimol & Peres, 2013).

The SAR is a key tool in conservation biogeography and, amongst other things, has been used to predict the number of extinctions resulting from habitat loss (e.g. Brooks, Pimm & Collar, 1997; Martins & Pereira, 2017), improve protected area design (e.g. Diamond, 1975), and predict the number of species occurring in large areas of natural habitat, such as a large expanse of tropical forest (Palmer, 1990; Rosenzweig, 1995; Plotkin et al., 2000; Desmet & Cowling, 2004; Santos et al., 2010; Smith, 2010; Basset et al., 2012; Gerstner, Dormann, Václavík, Kreft & Seppelt, 2014; Kunin et al., 2018). In regards to the latter, the ability to extrapolate the SAR to accurately predict the number of species occurring in large areas is of significant importance given the logistical and financial constraints involved in sampling over large spatial scales (Basset et al., 2012; Kunin et al., 2018). Typically, predicting richness at large spatial scales using the SAR is achieved by using the power model to predict the richness of an area (e.g. a large island, biome or region), either by using a set *z* value (generally around 0.25; Rosenzweig, 1995) or by estimating *z* from empirical data. However, as previously outlined, the power model may not always provide the best characterisation of the ISAR in empirical systems, and thus previous extrapolation studies based solely on the power model may have generated inaccurate predictions (this is true for any individual ISAR model). For example, Dengler (2009) compared the extrapolation ability of 12 ISAR models (in fact 25 models were compared as the same model was fitted using log-transformed and untransformed data; one model was applied using three different transformations) to accurately predict richness on large islands using six island archipelago datasets, and found that the mean rank of the power model was only 11th out of 25. Figure 1 provides a further illustration of this issue. Here, we have simulated eight islands of varying size (1, 3, 7, 14, 17, 22, 26, and 30; undefined units) that support reasonable numbers of species (3, 7, 14, 18, 20, 23, 24, and 25). We then fit five ISAR models (linear, logistic, negative exponential, power and Weibull3; see Table 1 for more details on these models) to these eight data points. Using these model fits, we estimated the number of species on an island of size 80 (grey dotted line in Fig.1) for each model and extrapolated each curve to its respective predicted value. It can be seen that the different models provide a range of predicted richness values for the hypothetical largest island.

An alternative extrapolation approach to simply using the power model is to use multi-model inference (MMI; Burnham & Anderson, 2002) and model averaging, whereby a larger number of *n* models is fitted to a set of islands, the models ranked according to some criterion (e.g. Akaike's information criterion, AIC; Burnham & Anderson, 2002) and the criterion values converted into model weights (i.e. the conditional probabilities for each of the *n* models; Wagenmakers & Farrell, 2004). The *n* models are then each used to predict the richness of a larger area and these predictions are multiplied by the respective model weights and summed to provide a multi-model averaged prediction (Burnham & Anderson, 2002; see Guilhaumon et al., 2008 for a SAR example).

A MMI approach is arguably much more robust as it provides a framework to deal with the model uncertainty observed in many SAR studies, and as Burnham & Anderson (2002, p. 198) note, such uncertainty can be much greater outside the range of the observed data. However, the effectiveness of the MMI framework in ISAR extrapolation is unknown, and

with the exception of the Dengler (2009) study that only analysed six island datasets, the question of model uncertainty in ISAR extrapolation has not been explored. As Dengler (2009, p.733) states, "although extrapolation of species richness beyond the largest plot size is one of the most frequent applications of SARs, there are only few and unsystematic approaches to testing which model function types are most suitable for this purpose."

It should be noted that using the ISAR is only one method for predicting the species richness of larger areas. For example, species accumulation curves, rarefaction methods and various extrapolation methods based on Hill numbers (Colwell & Coddington, 1994; Hsieh, Ma & Chao, 2016) are also widely used. However, many of these approaches require abundance data rather than incidence (i.e. presence-absence) data, although alternative methods are available for incidence data (see Hsieh et al., 2016). Incidence data are commonly available from biogeographical studies (e.g. Triantis et al., 2012; Matthews, Guilhaumon et al., 2016), which likely explains why the ISAR (which only requires incidence data) has often been used in extrapolation exercises (Dengler, 2009).

In this study, we use a set of 120 habitat island datasets to compare the accuracy of species richness extrapolation predictions using the power model with predictions using a model averaging approach based on twenty ISAR models. As such, our study goes beyond previous ISAR meta-analyses (e.g. Triantis et al., 2012; Matthews, Guilhaumon et al., 2016), which were focused on ISAR model goodness-of-fit evaluation, to explore ISAR model extrapolation capability. We focus on habitat islands rather than true islands (see Whittaker & Fernández-Palacios, 2007) as many applied SAR studies are focused on fragmented and forested terrestrial landscapes (e.g. Hubbell et al., 2008; Hanski, Zurita, Bellocq & Rybicki, 2013; Matthews, Cottee-Jones & Whittaker, 2014). We hypothesise that, due to the high degree of model uncertainty observed in many ISAR studies, the MMI framework will generate more accurate extrapolation predictions than the use of the power model on its own. The results of this analysis will provide useful information to guide future applications of ISAR extrapolation in conservation biogeography studies.

MATERIALS AND METHODS

Data collection

We took a subset of the habitat island datasets collected by Matthews, Cottee-Jones & Whittaker (2015) and Matthews, Guilhaumon et al. (2016). Habitat islands are defined as discrete habitat patches surrounded by contrasting matrix habitat. However, as in Matthews, Guilhaumon et al. (2016), we also included a small number of datasets consisting of protected areas for which the contrast between the matrix and the island was not so pronounced, and we included a few datasets of fragments within an aquatic matrix (e.g. rain forest fragment systems created by the construction of a reservoir), as the dominant assembly processes are considered to be more similar to those in habitat islands *sensu stricto* than oceanic islands (cf. Matthews et al., 2015). The original criteria for dataset collection (see Matthews, Guilhaumon et al., 2016) were: 1) the area and richness of each island were provided; 2) there

was no overlap between accepted datasets (data for different taxa within the same study system were accepted); and 3) there were at least four habitat islands. For the present study, we used datasets with at least eight islands and for which we could both successfully fit the power model (i.e. the model fit converged) and construct a multi-model averaged ISAR curve (i.e. at least two ISAR models could be successfully fitted to the dataset). We also manually (i.e. no explicit scale threshold was applied) filtered out datasets that were focused at very small spatial scales (e.g. insects on rose bushes or small experimental grassland plots) as these are not the spatial scale at which ISAR extrapolation is typically undertaken.

A total of 120 habitat island datasets were used, comprising 80 vertebrate, 21 plant, and 19 invertebrate datasets (Table S1 in Appendix S1 provides a summary of the datasets, and the source paper references are provided in Appendix S1).

Extrapolating the ISAR

To test the extrapolation ability of the various methods, we used the approach of Dengler (2009) whereby, for each dataset, we removed the largest island and all islands within a certain size threshold (*th*) relative to the largest island. For example, if the largest island was 100 ha and *th* was 0.5, we removed all islands larger than 50 ha. The new version of the dataset with the largest islands removed is referred to herein as the 'filtered dataset'. Removing the largest islands from each dataset allowed us to use the model fits to the filtered subset of islands to extrapolate and predict richness on larger islands for which we know the number of species. The value of *th* used in the main analyses was 0.5, although we experimented with different values as a sensitivity analysis (discussed below). For each filtered dataset, we then fitted the power (non-linear) ISAR model using non-linear regression and the 'sars' R package (version 1.1.1; Matthews, Triantis, Whittaker & Guilhaumon, 2019). With the exception of a model convergence check, the power model was fitted to a dataset regardless of the results of any model validation checks (the validity of this was tested as part of a sensitivity test, outlined below). A multi-model averaged ISAR curve was then fitted to the filtered dataset using the 'sar_average' function in the 'sars' R package. We attempted to fit twenty ISAR models (Table 1). A model was excluded if: 1) the model fitting process did not converge, 2) the model fit generated negative predicted values, 3) the residuals of the model fit were not normally distributed (using a Shapiro-Wilks test for normality), or 4) the residuals of the model fit were not homogeneous (assessed by correlating the residuals with the fitted values). All of these checks were undertaken using the 'sar_average' function (see Matthews et al., 2019). The remaining model fits were used to generate a multi-model averaged ISAR curve using AIC corrected for small sample size (AIC_c; Burnham & Anderson, 2002).

For each dataset, we followed the extrapolation procedure outlined in the introduction where we used the power model fit and the multi-model averaged curve to predict the species richness of the largest island in the original dataset (i.e. the largest of the islands that had been removed; see Dengler, 2009). In regards to the multi-model averaged curve, this worked by taking the multi-model fit object, using each of the individual model fits to predict the richness of the largest island, and multiplying these predictions by the respective AIC_c

weights. As AIC_c was used, for datasets where the filtered dataset had only six islands (7 cases when *th* = 0.5) it was not possible to calculate AIC_c for the 4 parameter ISAR models. Thus, the model weight was set to zero and the model fit had no bearing on the extrapolation prediction. As there was no functionality to undertake these extrapolations in the 'sars' R package, we wrote a new function to achieve this. The new function, 'sar_pred', takes two arguments (fit and area) and extrapolates the 'fit' object to predict the richness on an island of size 'area'. The 'fit' argument can be an individual SAR model fit (e.g. the power model) or a multi-model SAR curve. The new function is available in version 1.1.2 of the 'sars' package which is currently on GitHub (txm676/sars) and will be uploaded to CRAN shortly.

To compare the predictions of the power model and the multi-model averaged curve for a given dataset, we used the log error of extrapolation (LEE) metric of Dengler (2009) that addresses extrapolation capability. LEE is simply the log of the model's predicted richness minus the log of the observed richness (following Dengler, 2009, log to the base 10 was used); thus, the closer the LEE value is to zero the more accurate the prediction, and a positive LEE value means the model has over predicted the observed richness and *vice versa.* LEE was calculated for both the power model prediction and the multi-model averaged curve prediction.

As an important part of model prediction is to generate an estimate of the error of a prediction (Burnham & Anderson, 2002), the confidence intervals around the predictions were calculated using bootstrapping (Davison & Hinkley, 1997). For each of the filtered datasets, the data points (i.e. an individual island area and richness value) were sampled with replacement until the bootstrap sample was the same size as the original filtered dataset. The power model and multi-model curve prediction process was then undertaken using this bootstrap sample and the predictions stored. For the multi-model curve, the same models that were successfully fitted in the construction of the multi-model curve fit to the filtered dataset were selected. We did not undertake residual checks (e.g. normality) here to ensure bootstrap samples could be created, but we did still exclude model fits with negative predicted values. This process was repeated 100 times for each dataset and a 95% confidence interval constructed. Occasionally it was not possible to fit some of the relevant models to a bootstrap sample, or the predicted value was negative; in these cases, the bootstrap sample was discarded.

The main comparison of interest was the power model with the multi-model averaged curve. However, we also re-ran the above analysis including the extrapolation predictions of the additional 19 individual ISAR models. For each dataset, an individual model extrapolation prediction was included in the comparison only if the fit of the model to the filtered dataset passed all of the model validation checks.

Modelling variation in prediction accuracy

To determine whether any dataset characteristics could explain variation in the LEE values for the power model predictions (LEE-POW), we used generalized additive models (GAMs; Gaussian family) within a model selection framework. GAMs were used as there was evident

non-linear relationships between the predictors and the response. We used LEE-POW as the response variable. It was not possible to use the LEE values from the multi-model averaged curve (LEE-MMI) as the values were highly skewed and the residuals of the resultant models did approximate a normal distribution. For predictor variables, for each dataset (here the filtered dataset was used) we calculated the area of the smallest and largest islands and the ratio between them (A_{min}, A_{max} and A_{scale}), the richness of the most species poor and species rich islands and the ratio between them (S_{min}, S_{max} and S_{scale}), and the number of islands (Ni). For each dataset, we also took the latitude (Lat.) of the dataset and the sampled taxon (i.e. vertebrate, invertebrate or plant) from Matthews, Guilhaumon et al. (2016). Multicollinearity between predictors was tested using variance inflation factors: A_{max} and S_{max} were removed due to high multicollinearity and the remaining variance inflation factors were all below three. All of the continuous predictors (with the exception of latitude) were log-transformed to induce normality. The continuous predictors were modelled as penalized regression splines and the GAMs were fitted using the 'mgcv' R package (Wood, 2011). Smoothing parameter estimation was calculated using the Generalized Cross Validation (GCV) criterion.

A full set of models given all possible combinations of predictors were fitted using the MuMIn R package (Bartoń, 2012), and models were compared using AIC_c. The model with the lowest AIC_c value was considered the best model, and all models with delta-AICc <= 2 units of the best model were considered as having a similar degree of support (Burnham & Anderson, 2002). Model fits were validated using histograms of the residuals and plots of the residuals vs. the fitted values; the residuals of the full and best model roughly approximated a normal distribution and there were no evident patterns in the residuals. The relative importance of each predictor was calculated by summing the AIC_c weights for all models in which a predictor was included (Giam & Olden, 2016).

To determine whether the relative fit of a model to the filtered dataset explained its extrapolation performance, for each of the twenty models we calculated the LEE values across all datasets. For each ISAR model separately, we then fitted a simple generalized additive regression model (Gaussian family) whereby the absolute LEE values were the response variable and the AIC_c weights were the predictor variable, modelled as a penalized regression spline. Due to multiple testing, the critical P-value used was Bonferroni corrected (i.e. 0.05 / 20 = 0.0025).

Sensitivity analyses

To ensure our results were robust to the assumptions made during the analyses, we undertook three sensitivity tests. First, we re-ran the extrapolation analysis using *th* values of 0.3 and 0.7 (i.e. removing all islands that were 30% or 70% the size of the largest island in the original dataset). Second, in the main analyses, to ensure we could always compare the prediction of the power model with the prediction of the multi-model averaged curve we fitted the power model to all datasets regardless of the results of any model validation checks (with the exception of model convergence; e.g. no normality of residuals check was undertaken). Thus, we re-ran the prediction analysis after filtering out all datasets where the power model fit failed any of the following validation checks: 1) the model fit generated negative predicted

values, 2) the residuals of the model fit were not normally distributed, 3) the residuals of the model fit were not homogeneous, or 4) the *z* parameter was not significant. Third, we re-ran the prediction analysis after removing the linear model from the multi-model averaged curve fitting process (i.e. fitting of only 19 models was attempted; see Table 1). The reason for this third check is that previous studies have found that the linear model tends to provide a better relative fit to datasets with smaller numbers of islands, whereas in larger datasets its relative performance declines (e.g. Matthews, Guilhaumon et al., 2016). As the removal of larger islands necessarily generates datasets with fewer numbers of islands, it is possible that the linear model might provide better fits to the filtered datasets which then leads to inaccurate predictions if the ISAR of the full dataset is not linear. All analyses were undertaken using R (version 3.5.2; R Core Team, 2017). Unless stated otherwise, an alpha level of 0.05 was used in all significance tests.

RESULTS

When a *th* value of 0.5 was used, the power model provided the best fit to the most (filtered) datasets (n = 29), followed by the linear model (n = 21), and then the Monod (n = 19) and logarithmic models (n = 16) (see Table 1), according to AIC_c.

The full results of the main extrapolation and prediction analysis are provided in Table S2 in Appendix S2. In contrast to our hypothesis, the power model provided the most accurate prediction of the richness of the largest island (i.e. the lowest absolute LEE value) in 69 cases (58%), with the multi-model averaged curve providing the more accurate prediction in the remaining 51 cases (42%). The median LEE value of the power model was 0.04 (95% quantiles = -0.32 and 0.31), whilst the median LEE value of the multi-model curve (LEE-MMI) was 0.03 (95% quantiles = -0.35 and 0.74). However, as LEE values could be both positive and negative, the median of the absolute LEE values provides a better summary of the extrapolation capability: the median of absolute LEE-POW values was 0.08 (95% quantiles = 0.01 and 0.34), whilst the median of absolute LEE-MMI values was 0.10 (95% quantiles = 0.01 and 0.74). Both the power model (61% of LEE-POW values were positive) and the multi-model averaged curve (60% of LEE-MMI values were positive) had a slightly greater tendency to over predict the observed richness. The confidence intervals were on average narrower for the power model predictions (median 95% confidence interval width = 18) than for the multi-model averaged curve predictions (median 95% confidence interval width = 78) (Table S3 in Appendix S2). The confidence intervals around the multi-model averaged curve predictions were sometimes very large (i.e. spanning multiple orders of magnitude; see Table S3).

When the extrapolation predictions from all 20 ISAR models were considered, in addition to the multi-model averaged curve, the power model provided the most accurate prediction of the richness of the largest island in 11 cases, with the multi-model averaged curve providing the most accurate prediction in five cases. The Extended Power 2 (see Table 1) model provided the best prediction the most times, with 12 cases (the results for all models are provided in Table 1).

The full GAM (i.e. the GAM with all predictors) had a lower AIC score (-121.9) than an equivalent standard linear regression model (-115.4); this provides additional justification for our use of GAMs. When LEE-POW was used as the response variable in a GAM model selection analysis, the best model contained A_{scale}, S_{scale}, Lat. and S_{min} (Table 2). A plot of the smoothers for these four variables is provided as Figure 2. The effective degrees of freedom of the smoothers for A_{scale} and Lat. were one, indicating that these smoothers were straight lines; increasing A_{scale} resulted in decreasing LEE-POW, while the opposite pattern was true for Lat (Fig. 2). The S_{scale} and S_{min} relationships were more complex (Fig. 2), but increasing S_{scale} resulted in an approximate increase in LEE-POW. However, there was a reasonable degree of model uncertainty as the best model had an AIC_c weight of only 0.20, and there were two additional models within 2 delta AIC_c units of the best model (Table 2). In addition, the adjusted R^2 value of the best model was low (0.20). A_{scale} (0.98), S_{scale} (0.95) and S_{min} (0.81) had quite high relative importance values, whilst the values for the remaining predictors were all lower (Table 2).

For 18 of the ISAR models, the relative fit of a model to the filtered dataset (i.e. the model's AIC_c weight) was a poor predictor of a model's extrapolation accuracy (measured using the LEE metric). In only two cases (for the Power Rosenzweig and Extended Power 1 models; see Table 1 for model descriptions) was the AIC_c weight a significant predictor of a model's absolute LEE value (Table S4 in Appendix S2).

The choice of *th* value did not change the overall qualitative results. The power model provided the more accurate prediction in 65 (54%) and 72 (61%; when a *th* value of 0.7 was used there was one dataset for which no models could be successfully fitted) cases when *th* values of 0.3 and 0.7 were used, respectively (see Table S5 & S6 in Appendix S2). In regards to the power model validation sensitivity test, there were 23 datasets for which the power model failed one of the validation checks. However, removing these 23 datasets and re-running the prediction analysis using the remaining 97 datasets did not change the overall qualitative results: the power model provided the most accurate prediction in 55 cases (57%). Finally, re-running the prediction analysis after excluding the linear model from the multi-model averaged curve resulted in a slight increase in the number of cases where the multi-model averaged curve provided the more accurate prediction (60 out of 120 cases), but the general picture remained the same.

DISCUSSION

Using 120 habitat island datasets, we compared the extrapolation capability of the power ISAR model with that of a multi-model averaged ISAR curve constructed using up to twenty ISAR models. In contrast to our hypothesis that the multi-model curve would produce more accurate species richness predictions, we found that the power model provided the more accurate prediction in a majority of cases.

Model averaging is not a panacea for ISAR extrapolation

It is rarely feasible to produce complete inventories of all species of a given taxonomic group at large spatial scales (e.g. in a large expanse of tropical forest or on very large islands;

Colwell & Coddington, 1994). The question of how to extrapolate from samples collected at relatively small scales to accurately predict richness over larger areas is therefore the subject of considerable research effort (Hsieh et al., 2016). There has been particular focus on the ISAR (in addition to SARs constructed using continuous habitat data) as it only requires incidence data; yet a statistically rigorous ISAR extrapolation method, required for accurate richness predictions, has proven elusive. The present study represents a formative step in the development of such a method.

Based on the results and arguments presented in many recent SAR studies and other model prediction exercises (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004; Guilhaumon et al., 2008, 2010; Triantis et al., 2012; Benchimol & Peres, 2013; Matthews, Guilhaumon et al., 2016), we hypothesised that a model averaging framework would result in more accurate extrapolation predictions than simply using the power model in isolation. Model averaging based on a set of competing candidate models has been proposed for improving predictions in cases where model uncertainty is prevalent (as seems to be the case with the ISAR; e.g. Triantis et al., 2012). For example, Burnham & Anderson (2002, p.150) state that "prediction is an ideal way to view model averaging, because each model in a set, regardless of its parameterisation, can be used to make a predicted value." However, in contrast to our hypothesis we found that the power model provided the best prediction in the most cases, regardless of which *th* value was used (0.3, 0.5 or 0.7). Although our study is the first comprehensive evaluation of model averaging in ISAR extrapolation, Mazel et al. (2014) found that the power model on its own provided similar results to a multi-model average curve when using SARs, and functional diversity- and phylogenetic diversity-area relationships, to select biodiversity hotspots. Thus, it may be that the power model is generally a more appropriate tool than multi-model averaged curves in many applied SAR contexts. More research is needed to examine the performance of multi-model averaging in other areas of applied SAR research, such as predicting the number of extinctions resulting from habitat loss.

Analysis of the raw LEE-POW and LEE-MMI values indicates that both the multi-model averaged curve and the power model had a slightly greater tendency for over-prediction of species richness. The tendency of the power model to overpredict richness has been previously documented (Palmer, 1990; Smith, 2010), but the performance of multi-model averaged ISAR curves when extrapolating richness has not been previously documented. A tendency for over prediction is arguably preferable as, in a conservation context, underprediction bias is likely to carry greater risk (for instance when forecasting the impacts of habitat loss).

In general, the multi-model curve predictions also exhibited wider confidence intervals (Table S3). In certain cases, these were very wide, reflecting the bootstrap procedure that we employed, which works by sampling islands (paired area and richness values) with replacement; this process can result in the same island being chosen multiple times, particularly in smaller datasets, resulting in some bootstrapped samples having unusually-shaped ISARs which do not bear much resemblance to the ISAR of the original sample/dataset. As the multi-model curve combines multiple ISAR models it is necessarily

more flexible, allowing it to more accurately characterise the form of the unusually-shaped bootstrapped ISARs, but which can then result in wayward extrapolation predictions (i.e. predictions of the largest island in the original dataset).

Interestingly, when the extrapolation predictions from all individual twenty ISAR models were compared the Extended Power 2 model (EPM2; Table 1) provided the most accurate predictions the most times (12 times vs. 11 times for the standard power model). The EPM2 model, which is a sigmoidal model, is from within the same family as the standard power model (i.e. it is defined by adding a single additional parameter to the standard power model; Tjørve, 2009). The greater flexibility that arises from an additional parameter necessarily means that the EPM2 model should explain more variation in richness than the standard power model (i.e. have a larger R^2); however, this does not mean the model should produce more accurate extrapolation predictions. For example, the Extended Power 1 model, which is also in the same model family as the power model, only provided the most accurate prediction 5 times. In addition, the other sigmoidal models generally performed poorly (Table 1). In contrast to Tjørve (2009), who postulated that extended power models may provide poor extrapolation predictions, these results call for greater assessment of extended power models in applied ISAR applications.

Why does the power model provide better predictions on average?

The rationale for the smaller confidence intervals around the extrapolation predictions of the power model described in the preceding paragraph also provides an explanation for why the multi-model curve provided less accurate predictions in a majority of cases more generally: the greater flexibility of the multi-model curve is also its downfall. Regardless of the shape of the ISAR of the full dataset, unless that shape is characterised by a linear model, the form of the filtered dataset will differ, often considerably, from that of the full dataset. One of the advantages of the MMI approach, if the model set contains a range of sensible models given the situation, is that it often provides a better fit to a set of data than any one model on its own (Burnham & Anderson, 2002). However, if the shape of the filtered dataset is not representative of that of the full dataset, this greater flexibility may be a negative feature. For example, the linear model has been shown to provide a better fit relatively speaking to datasets with few, relatively smaller, islands (Matthews, Guilhaumon et al., 2016). Thus, it can be assumed that the relative performance of the linear model is better for the filtered datasets than for the full datasets; this better performance means it will have a larger information criterion weight and thus a stronger influence on the multi-model curve. However, if the full dataset is actually even just somewhat convex the multi-model curve (with its linear element) will not provide an accurate extrapolation prediction. In addition, it may be that habitat island datasets contain substantial amounts of noise due to the role of factors other than area (e.g. human disturbance; Benchimol & Peres, 2013). These factors, which may have a greater relative effect in small fragments (Matthews et al., 2014), may result in "messy" ISAR datasets. The more complex models have greater flexibility to fit this noise, resulting in poor extrapolation behaviour. For example, in a small number of cases, the largest fragment in the filtered dataset had lower richness than some of the smaller fragments,

resulting in some of the more complex models predicting decreasing richness with increasing area and thus predicting negative richness when extrapolated!

Explaining variation in extrapolation capability of the power model across datasets

Our generalized additive model selection analysis indicated that the most important variables in driving variation in LEE-POW across datasets were A_{scale}, S_{scale}, S_{min} and Lat. (Table 2), with A_{scale}, S_{scale} and S_{min} in particular having relative importance values greater than 0.80. It should be noted that the amount of variation in LEE-POW explained by the best model was relatively low (adjusted $R^2 = 0.20$). In the best model, the effect of A_{scale} on LEE-POW was linear and negative, whilst the effect of S_{scale} was non-linear but broadly positive and convex (Fig. 2). These results indicate that increasing A_{scale} results in lower LEE-POW values while, in contrast, increasing S_{scale} results in larger LEE-POW values, although there is a flattening out of this latter relationship at larger values of S_{scale} (Fig. 2). The negative effect of A_{scale} on LEE-POW values is logical because the full convex shape of the empirical ISAR may only become apparent when a large range of island sizes is studied (Martin, 1981; Matthews, Guilhaumon et al., 2016); for a smaller range of island areas the relative performance of the linear model is conversely greater. Thus, if A_{scale} is small and, in particular, there are no relatively large fragments within the dataset, the ISAR is less likely to be characterised by a power model (and more likely by a linear model) and attendant extrapolation predictions are likely to over-predict the true richness value. The positive effect of S_{scale} is more surprising, as one would expect the range in species richness in a dataset to scale positively with the range in island area. Indeed, A_{scale} and S_{scale} were significantly, albeit weakly, positively correlated (Spearman's rho = 0.38; $P < 0.001$). We speculate that S_{scale} co-varies with another variable that was not included in our analysis, such as sample completeness (Hsieh et al., 2016). For example, if S_{scale} is related to the number of species across all fragments (information that is not available from ISAR datasets) and more species-rich taxa are more likely to have been under-sampled, particularly in the larger fragments, then the effect of S_{scale} may in fact be evidence of a sampling artefact. Further research is needed to explore this possibility.

We also found that, generally speaking, a model's relative fit to the filtered dataset provided a poor predictor of that model's extrapolation accuracy. This further complicates providing general guidelines for extrapolation as it rules out simply selecting the best fitting model when undertaking ISAR extrapolation.

Conclusions

Our findings show that multi-model averaging is unlikely to provide a universally suitable method for ISAR extrapolation, even though there is a large amount of model uncertainty (e.g. see the mean AIC_c weights of each model in Table 1). Taking the specific characteristics of the studied dataset into account (e.g. island size range, species richness range) could lead to more informed ISAR model selection, though this requires further investigation. However, the relevance of our results is likely to be restricted to the spatial scale of the analysed datasets. Although some of our datasets contain very large islands (largest island across all

datasets = 19,604 km^2), the median island size is much smaller (0.09 km^2), and our results may thus not be transferable to i) scenarios requiring the ISAR to be extrapolated to very large areas (e.g. biotic regions or provinces; Rosenzweig, 1995; Gerstner et al., 2014), or ii) other types of SARs (e.g. species accumulation curves; Bassett et al., 2012; Kunin et al., 2018). It is also possible that habitat island datasets are particularly noisy and that we may find different results when looking at true islands, for example.

Although the power model provided more accurate predictions in a majority of cases, it is hard to advocate blanket use of the power model in future ISAR extrapolation analyses, as in approximately 40% of cases the multi-model averaged curve provided a better prediction. Depending on the aim of the study, a comparative selection of techniques (e.g. multiple individual ISAR models and the multi-model averaged curve) may be useful, yielding a range of predictions with confidence intervals that can be assessed together. In situations where a single point estimate is required, our results would support judicious use of the power model. However, further research at larger spatial scales is urgently needed to validate these recommendations for ISAR extrapolation in a wider context.

ACKNOWLEDGEMENTS

The recently published 'sars' R package, which was used to run the analyses in the paper, was written in collaboration with François Guilhaumon. François Rigal provided modelling advice. Two anonymous reviewers provided comments that improved the paper.

DATA ACCESSIBILITY

All datasets are publicly available and the full source citations are provided in the Supporting Information.

REFERENCES

Arrhenius, O. (1921). Species and area. *Journal of Ecology*, 9, 95-99.

Bartoń, K. (2012). MuMIn: multi-model inference (R package version 1.40.4). Retrieved from https://cran.r-project.org/web/packages/MuMIn/index.html

Basset, Y., Cizek, L., Cuénoud, P., Didham, R. K., Guilhaumon, F., Missa, O., . . . Leponce, M. (2012). Arthropod diversity in a tropical forest. *Science,* 338, 1481-1484.

Benchimol, M., & Peres, C. A. (2013). Anthropogenic modulators of species–area relationships in Neotropical primates: a continental-scale analysis of fragmented forest landscapes. *Diversity and Distributions,* 19, 1339-1352.

528 Brooks, T.M., Pimm, S.L. & Collar, N.J. (1997). Deforestation predicts the number of
529 threatened birds in insular Southeast Asia. *Conservation Biology*, 11, 382-394.

530 Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: a*
531 *practical information-theoretic approach* (2nd ed.). New-York: Springer.

532 Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through
533 extrapolation. *Philosophical Transactions: Biological Sciences*, 345, 101-118.

534 Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*.
535 Cambridge: Cambridge University Press.

536 Dengler, J. (2009). Which function describes the species–area relationship best? A review
537 and empirical evaluation. *Journal of Biogeography*, 36, 728-744.

538 Desmet, P., & Cowling, R. (2004). Using the species-area relationship to set baseline targets
539 for conservation. *Ecology and Society*, 9, 11-33.

540 Diamond, J.M. (1975) The island dilemma: lessons of modern biogeographic studies for the
541 design of natural reserves. *Biological Conservation*, 7**,** 129-146.

542 Gerstner, K., Dormann, C. F., Václavík, T., Kreft, H., & Seppelt, R. (2014). Accounting for
543 geographical variation in species–area relationships improves the prediction of plant
544 species richness at the global scale. *Journal of Biogeography*, 41, 261-273.

545 Giam, X., & Olden, J. D. (2016). Quantifying variable importance in a multimodel inference
546 framework. *Methods in Ecology and Evolution*, 7, 388-397.

547 Guilhaumon, F., Gimenez, O., Gaston, K. J., & Mouillot, D. (2008). Taxonomic and regional
548 uncertainty in species-area relationships and the identification of richness hotspots.
549 *Proceedings of the National Academy of Sciences USA*, 105, 15458-15463.

550 Guilhaumon, F., Mouillot, D., & Gimenez, O. (2010). mmSAR: an R-package for multimodel
551 species–area relationship inference. *Ecography*, 33, 420-424.

552 Hanski, I., Zurita, G. A., Bellocq, M. I., & Rybicki, J. (2013). Species–fragmented area
553 relationship. *Proceedings of the National Academy of Sciences USA*, 110, 12715-
554 12720.

555 He, F. & Legendre, P. (1996). On species-area relations. *The American Naturalist*, 148, 719-
556 737.

557 Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: an R package for rarefaction and
558 extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution*,
559 7, 1451-1456.

560 Hubbell, S. P., He, F., Condit, R., Borda-de-Água, L., Kellner, J., & ter Steege, H. (2008).
561 How many tree species are there in the Amazon and how many of them will go
562 extinct? *Proceedings of the National Academy of Sciences USA*, 105, 11498-11504.

Kunin, W. E., Harte, J., He, F., Hui, C., Jobe, R. T., Ostling, A., . . . Varma, V. (2018). Upscaling biodiversity: estimating the species–area relationship from small samples. *Ecological Monographs,* 88, 170-187.

Lomolino, M. V. (2000). Ecology's most general, yet protean pattern: the species-area relationship. *Journal of Biogeography,* 27, 17-26.

Martin, T.E. (1981). Species-area slopes and coefficients: a caution on their interpretation. *The American Naturalist,* 118**,** 823-837.

Martins, I. S., & Pereira, H. M. (2017). Improving extinction projections across scales and habitats using the countryside species-area relationship. *Scientific Reports,* 7, 12899.

Matthews, T. J., Cottee-Jones, H. E., & Whittaker, R. J. (2014). Habitat fragmentation and the species–area relationship: a focus on total species richness obscures the impact of habitat loss on habitat specialists. *Diversity and Distributions,* 20, 1136-1146.

Matthews, T.J., Cottee-Jones, H.E.W. & Whittaker, R.J. (2015). Quantifying and interpreting nestedness in habitat islands: a synthetic analysis of multiple datasets. *Diversity and Distributions,* 21**,** 392-404.

Matthews, T. J., Guilhaumon, F., Triantis, K. A., Borregaard, M. K., & Whittaker, R. J. (2016). On the form of species–area relationships in habitat islands and true islands. *Global Ecology and Biogeography,* 25, 847–858.

Matthews, T. J., Triantis, K. A., Rigal, F., Borregaard, M. K., Guilhaumon, F., & Whittaker, R. J. (2016). Island species–area relationships and species accumulation curves are not equivalent: an analysis of habitat island datasets. *Global Ecology and Biogeography,* 25, 607-618.

Matthews, T. J., Triantis, K. A., Whittaker, R. J., & Guilhaumon, F. (2019). sars: an R package for fitting, evaluating and comparing species–area relationship models. *Ecography, In press.*

Mazel, F., Guilhaumon, F., Mouquet, N., Devictor, V., Gravel, D., Renaud, J., . . . Thuiller, W. (2014). Multifaceted diversity–area relationships reveal global hotspots of mammalian species, trait and lineage diversity. *Global Ecology and Biogeography,* 23, 836-847.

Palmer, M. W. (1990). The estimation of species richness by extrapolation. *Ecology,* 71, 1195-1198.

Plotkin, J. B., Potts, M. D., Yu, D. W., Bunyavejchewin, S., Condit, R., Foster, R., . . . Ashton, P. S. (2000). Predicting species diversity in tropical forests. *Proceedings of the National Academy of Sciences USA,* 97, 10850-10854.

R Core Team. (2017). R: a language and environment for statistical computing (Version 3.5.1). Vienna, Austria: R foundation for statistical computing. Retrieved from https://www.R-project.org/

Rosenzweig, M. L. (1995). *Species diversity in space and time*. Cambridge: Cambridge University Press.

Santos, A.M.C., Whittaker, R.J., Triantis, K.A., Borges, P.A.V., Jones, O.R., Quicke, D.L.J. & Hortal, J. (2010). Are species–area relationships from entire archipelagos congruent with those of their constituent islands? *Global Ecology and Biogeography,* 19**,** 527-540.

Scheiner, S. M. (2003). Six types of species-area curves. *Global Ecology and Biogeography*, 12, 441-447.

Smith, A. B. (2010). Caution with curves: caveats for using the species-area relationship in conservation. *Biological Conservation,* 143, 555-564.

Tjørve, E. (2003). Shapes and functions of species–area curves: a review of possible models. *Journal of Biogeography,* 30, 827-835.

Tjørve, E. (2009). Shapes and functions of species–area curves (II): a review of new models and parameterizations. *Journal of Biogeography*, 36, 1435-1445.

Tjørve, E., & Tjørve, K. M. C. (2017). Species-area relationship. *eLS* (*Encyclopedia of Life Sciences Online).* Chichester: John Wiley & Sons.

Triantis, K. A., Guilhaumon, F., & Whittaker, R. J. (2012). The island species–area relationship: biology and statistics. *Journal of Biogeography,* 39, 215-231.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review,* 11, 192-196.

Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: ecology, evolution, and conservation* (2nd ed.). Oxford: Oxford University Press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 73, 3-36.

BIOSKETCH

Tom Matthews is a macroecologist and biogeographer at the University of Birmingham, UK. He is interested in the application of macroecological methods to global environmental change questions, and his previous work has focused on the impacts of habitat fragmentation and the form of the species–area relationship in fragmented landscapes.

Thomas Aspin is a disturbance ecologist affiliated with the University of Birmingham. His research broadly centres on the interface of disturbance ecology, macroecology and conservation ecology, with particular emphasis on climate change and habitat loss.

Author Contributions: TJM designed the study and collected the data; TJM ran the analyses with input from TWHA; TJM and TWHA wrote the paper.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

643 **TABLES**
644
645 **Table 1** The twenty models that were fitted to generate the multi-model averaged ISAR
646 curve. The model shape is the general model shape, as in Triantis et al. (2012); the observed
647 shape can deviate from the general model shape in cases when fitting certain models. For the
648 model equation, A = sample area, and d, c, z and f are free parameters. Each equation is
649 calculating the number of species. Mean weight is the mean AIC_c weight for a given model
650 across all fits to the filtered datasets (excluding non-satisfactory fits). Best fit corresponds the
651 number of times a model provided the best fit to a filtered dataset (i.e. had the lowest AIC_c
652 value). Best prediction corresponds to the number of times a model provided the best
653 extrapolated prediction in the all model comparison; these values do not sum to 120 (the
654 number of datasets) as the multi-model averaged curve provided the best extrapolation
655 prediction in five cases.
656

Model	No. parameters	Model shape	Equation	Mean weight	Best fit	Best Prediction
Asymptotic	3	Convex	d - c*z^A	0.04	0	6
Beta-P	4	Sigmoid	d*(1-(1+(A/c)^z)^-f)	<0.01	0	4
Chapman–Richards	3	Sigmoid	d * (1 - exp(-z*A)^c)	0.01	0	6
Logarithmic	2	Convex	c+z*log(A)	0.14	16	10
Extended Power 1	3	Convex/Sigmoid	c*A^(z*A^-d)	0.04	0	5
Extended Power 2	3	Sigmoid	c*A^(z-(d/A))	0.03	1	12
Gompertz	3	Sigmoid	d*exp(-exp(-z*(A-c)))	0.04	2	4
Kobayashi	2	Convex	c*log(1 + A/z)	0.15	13	5
Linear	2	Linear	c + z*A	0.12	21	9
Logistic	3	Sigmoid	c/(f + A^(-z))	0.03	0	7
Monod	2	Convex	d/(1+c*A^(-1))	0.10	19	7
Morgan–Mercer–Flodin	3	Sigmoid	d/(1+c*A^(-z))	0.03	0	1
Negative Exponential	2	Convex	d*(1-exp(-z*A))	0.10	11	4
Persistence Function 1	3	Convex	c*A^z * exp(-d*A)	0.03	2	2
Persistence Function 2	3	Sigmoid	c*A^z * exp(-d/A)	0.04	2	6
Power	2	Convex	c*A^z	0.16	29	11
Power Rosenzweig	3	Convex	f + c*A^z	0.03	1	6
Rational	3	Convex	(c + z*A)/(1+d*A)	0.03	1	3
Weibull-3	3	Sigmoid	d*(1 - exp(-c*A^z))	0.04	1	4
Weibull-4	4	Sigmoid	d * (1 - exp(-c*A^z))^f	0.01	1	3

657
658
659
660
661
662
663

Table 2 The results of the generalized additive model selection. The response variable was the LEE values from 120 habitat island datasets for the power model curve (see the main text), which provides an assessment of the extrapolation accuracy of the power ISAR model. The predictor variables were the smallest island area in a dataset (A_{min}) and the ratio between the largest and the smallest island area (A_{scale}), the same two variables but for species richness (S_{min} and S_{scale}), the number of islands in a dataset (Ni), the latitude of the dataset (Lat.) and the taxon sampled (Taxon). A_{min}, A_{scale}, Lat, Ni, S_{min} and S_{scale} were all modelled as 'penalized regression splines', while taxon was modelled as a standard linear variable (as it was categorical). A '+' indicates that a variable was included within a model. Models were ranked using AIC_c and all models with delta AIC_c values less than two are shown. The AIC_c weight of each model is also provided. The relative importance (RI) of each predictor is shown on the bottom row.

Model	A_{min}	A_{scale}	Lat.	Ni	S_{min}	S_{scale}	Taxon	Delta	Weight
1	-	+	+	-	+	+	-	0.00	0.20
2	-	+	-	-	+	+	-	0.48	0.16
3	+	+	+	-	+	+	-	1.89	0.08
RI	0.27	0.98	0.55	0.25	0.81	0.95	0.15		

FIGURES

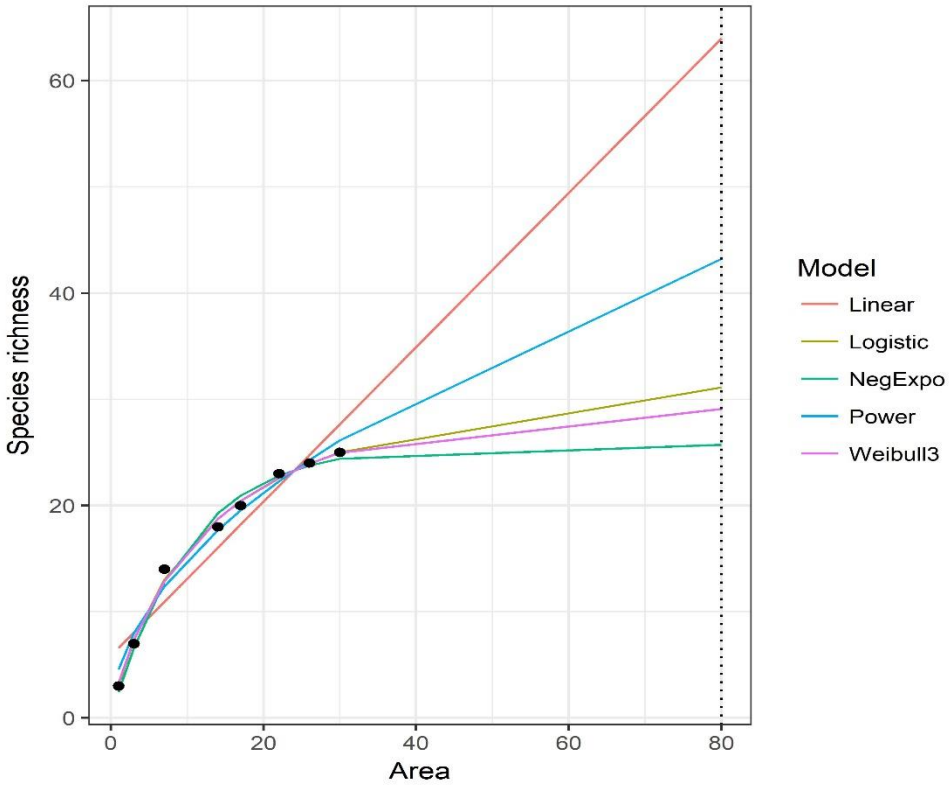
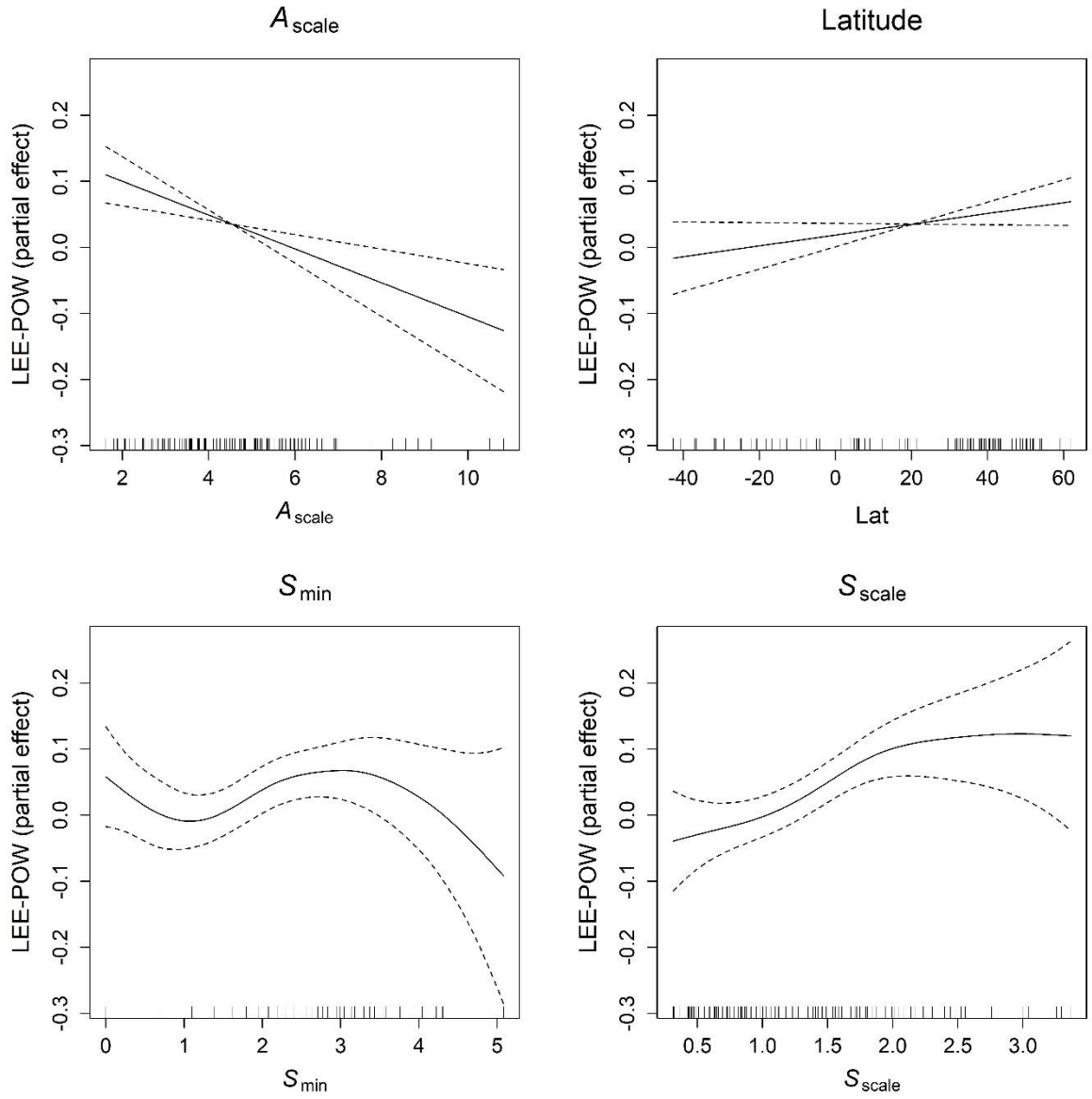


Figure 1 The varying species richness predictions of five ISAR models. Each of the five models (see Table 1) was fitted to a simulated archipelago consisting of eight islands of varying size (1, 3, 7, 14, 17, 22, 26, and 30; undefined units) and richness (3, 7, 14, 18, 20, 23, 24, and 25). These model fits were then used to predict the richness of an island of size 80 (grey dotted line).



687

688 **Figure 2** Fitted smoothers from the best fit generalized additive model showing the partial
689 effects of A_{scale}, Latitude, S_{min} and S_{scale} on the LEE-POW values. The fitted values have been
690 shifted in each plot by adding the model intercept (0.04) value (using the shift argument in
691 the plot.gam R function). The effective degrees of freedom for each smoother are: A_{scale}
692 (1.00), Latitude (1.00), S_{min} (3.53) and S_{scale} (2.70). The dashed lines represent the standard
693 error curves (two SE above and below). Each LEE-POW value relates to the accuracy of a
694 prediction of the number of species on a habitat island using the power model. For each of
695 120 habitat island datasets, the largest island and all islands larger than half the size of the
696 largest island were removed and the power model fitted to the filtered dataset and
697 extrapolated.

698

699